Application of Data Warehouse and Data Mining in Construction Management

Jianping ZHANG[1] (zhangjp@tsinghua.edu.cn)

Tianyi MA[1] (matianyi97@mails.tsinghua.edu.cn)

Qiping SHEN[2] (bsqpshen@inet.polyu.edu.hk)

[1]Department of Civil Engineering, Tsinghua University, Beijing, China

[2]Department of Building and Real Estate, The Hong Kong Polytechnic University,

Hunghom, Kowloon, Hong Kong, China

Summary

All construction project are constrained by their schedules, budgets and specifications, and safety and environmental regulations. These constraints made construction management more complex and difficult. At the same time, many historical data that can support the decisions in the future are kept in construction enterprises,. To use the historical data effectively and efficiently, it is essential to apply the data warehouse and data mining technologies. This paper introduces a research which aims to develop a data warehouse system according to the requirements of construction enterprises and use data mining technology to learn useful information and knowledge from the data warehouse system. The design, the development and the application of this system are detailedly introduced in this paper.

1 Introduction

In China, many construction projects have large scales and complex structures. The management of these projects involves many labors, materials, machineries, equipments and finance. At the same time, the enterprise must control the schedule, quality, site safety and cost of projects in the complex natural and social environment. All of these factors bring enormous complexity. For an enterprise to take on several projects synchronously brings them highly request on the ability of management and decision-making. At present, the management method and system of many construction enterprises in China is still behindhand. Their management and decision-making is also depending on the experience and ability of the managers. Because China has entered WTO, construction enterprises have to face fierce competition from many international enterprises. Chinese construction enterprises must apply the latest information technology in the management of construction projects, aiming at the enterprise's information management system. With the help of this management system, managers can analyze the enterprise's working state and forecast its development trends.

To improve the decision-making ability of an enterprise, information technology must be applied to each step in the management of enterprise. It can realize the automation of collection,

management, analysis, and utilisation of information. During the construction management process, enterprises can accumulate valuable historic data. The key problem is however, how to organize and analyze these data, so as to obtain internal principles and help managers make decisions. The development of data-warehouse (DW) and data-mining (DM) technology offers possible solutions to these problems. DW has been applied for several years in some industries, such as finance, insurance, telecommunication. But in construction, the application of information techonology is just starting. The DW and DM technology has not systemticly applied in construction enterprises.

This paper introduces a research which is established in construction enterprises' requirement, and use DW and DM technology to build a DW and DM system of construction enterprises. The primary objective of this system is to assistant construction enterprise by making decisions during the management of projects.

2 Data Warehouse and Data Mining

Inmon (1993), in his landmark work *Building the Data Warehouse*, offers the following definition of data warehouse: "a data warehouse is a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management's decision making process."

An enterprise's DW is a solution and mainly a process of establishment. It summarizes the enterprise's requirement, analyzes the enterprise's operation system, and forms the enterprise's DW system using the data of OLTP system by cleaning up, collecting and loading. The data stored in the DW are enterprise's historic data. The accumulation of these historic data needs a long time, and the development of DW is a process of ameliorate constantly. The DW's revolution will become better and better along with use and the information, obtained from the DW system, is more credible. So DW is a kind of technology, but on the other hand, it is an idea, a solution of problems, a system of reframing data.

DM, on the other hand, is referred to as Knowledge Discover Database (KDD). It is about the techniques for the extraction of useful information and implicit knowledge from large databases or other data. These information and knowledge may never be known by people. DM is a process of discovering implicit pattern and relationship in a huge database. It can discover description and prediction information. The two major tasks in DM are description and prediction. Description can depict data's generic characters and summarize implicit disciplinarian in these data. Prediction can forecast much valuable information or conclusion according to some actual data and some necessary condition.

DM can be sorted into class/concept description, association analysis, classification, clustering, outlier mining, evolution analysis, and so on (Han, 2001). The foundation of DM is the data, hence it must be related to the technologies of database, DW, OLTP, OLAP, etc. Besides data, the DM algorithms are another important aspect of its successful application. There are conceptive differences between Mining Models and Mining Algorithms. Mining models are referred to as physical structures of data subset, compiled by mining algorithms, and involve the description of original dataset. Moreover DM algorithms are mathematic and statistic methods, which transform original data instances into mining models (Seidman, 2002).

3 The Data-Warehouse System of Construction Enterprises

The development of a DW system is a gradual and circular process considering the multi-goal and the complexity of the system. Generally, requirements of a DW system cannot be clearly defined by construction enterprises at the beginning of a DW project, since enterprises do not know exactly what a DW system is. The developers are supposed to help consumers understand by building a DW model, and providing tools of building and managing the DW. Once the concept of a DW is clear, the enterprise can submit its actual requirements to the developers continually. According to these new requirements, the DW system is refined. After several cycles, the DW will meet the consumers' requirements on the whole. The procedure of establishing an enterprise DW system is shown as Figure 1.

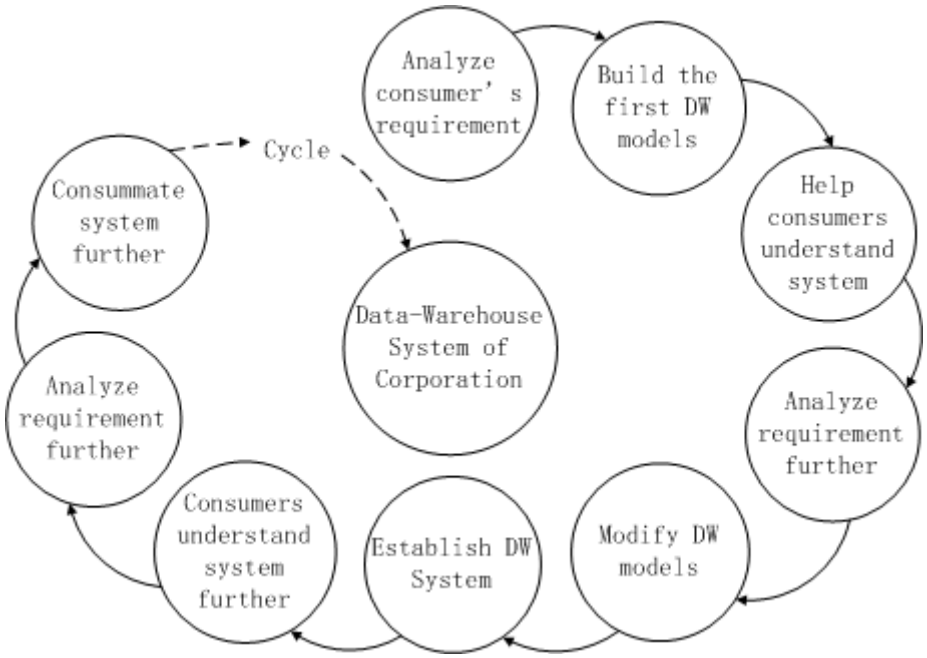

Figure 1. The establishment procedure of a DW system

3.1 Design of the Data-Warehouse Models

According to the investigation on construction enterprises, a series of subjects are defined in the construction enterprise DW, including human resources, material, machine, schedule, quality, safety, cost, etc.

The subject structure of DW model includes star schema and snowflake schema. A star schema is a specific type of database design used to support analytical processing, which includes a specific set of de-normalized tables. A star schema contains two types of tables: fact tables and dimension tables. Fact tables contain the quantitative or factual data about a construction management entity. Dimension tables are smaller and hold descriptive data that reflect the dimensions of an entity (Chau K.W.and Ying Cao, 2002). A typical star schema model of material inventory is shown as Figure 2. This system's DW is mainly designed with star schemas, but there are also several snowflake schemas.

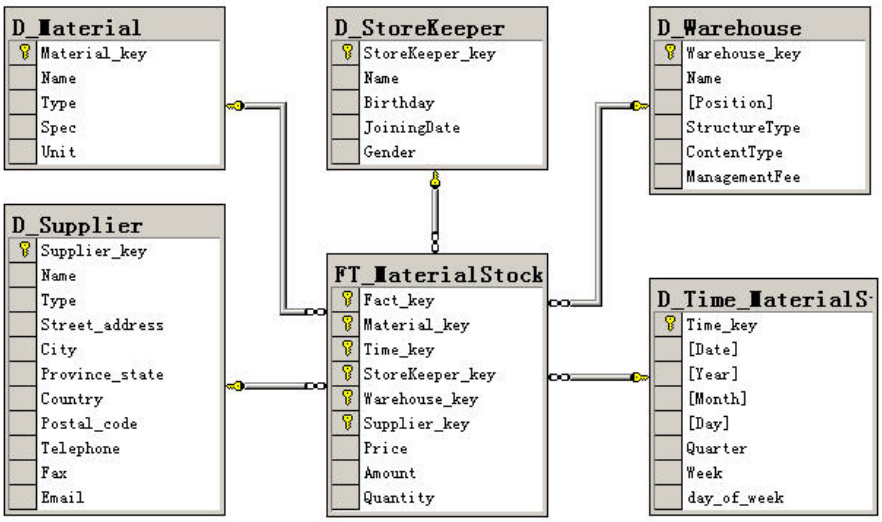

Figure 2 The star schema of material inventory

3.2 Development of DW system

To design and develop components both in server and client are required in the development of DW system based on DW models. These components mainly consist:

- ♦ Tools and components for data extraction, cleaning, intergation and population on server;
- ♦ Components for data cube building and management on server;
- ♦ Tools for data query on clients;

Microsoft SQL Server 2000 is chosen as the database platform of the system in this project, while the tools and interface shipped with SQL Server are used in programming.

3.2.1 The development of tools and components for data extraction, cleaning, intergation and population on server

The component object models of Data Transformation Service (DTS) offered by SQL Server are used to develop these components. System can extract data from distributed OLTP databases of data files and load DW with data after clearing and integration. For the tasks that DTS can't achieve, some ActiveX controls were developed. These ActiveX controls can complete those special tasks. After completing these components, system administrator must termly load the DW with those data generated by the management information system (MIS) and data file of construction projects in progress. With the help of components offered by this system, administrators can complete these tasks easily. In this system, there are three primary data sources: CMS (Chau K.W.and Ying Cao, 2002), 4D-GCPSU (J.P. Zhang and H.J. Wang, 2002) and standard data files of construction projects (text, MS Word and MS Excel files). There are corresponding components of data extraction in allusion to each of these three data sources.

3.2.2 The development of components for data-cubes' building and management on server

After DW has been loaded with data, data in every subject model were transformed to Data Cube, this is a physical form adapted to the developing tools of clients. This process include two steps: establishing cubes' logical structure and determining the physical structure of logical structure. Usually, the amount of data in a data warehouse is every large, so the data in these cubes has been aggregated by pre-calculations. These aggregated cubes can accelerate speed of query and reduce response time.

3.2.3 Tools for data query on clients

The tools for data query in clients are developed by using Microsoft Visual Basic. Using ADOMD, MDX and OLAP service, our system supports aggregated data or detail data. With this tool, users can get useful and aggregated data by drilling-up, drilling-down, pivoting. First, user submits a query request for a data cube, and data meeting to query will be down-loaded to client in aggregated forms. Then user can pivot dimensions to different axis or change the order of dimensions in the same axis, click on axis to drill-down or dill up data, and select special members' value in every level of dimensions in order to get some specific data set. With these data, user can select some data in the grid to form all kinds of analysis and statistic graphs. The selected data region may be continuous or discrete. Figure 3 illustrates the interface of the system with an example of *material use* cube.

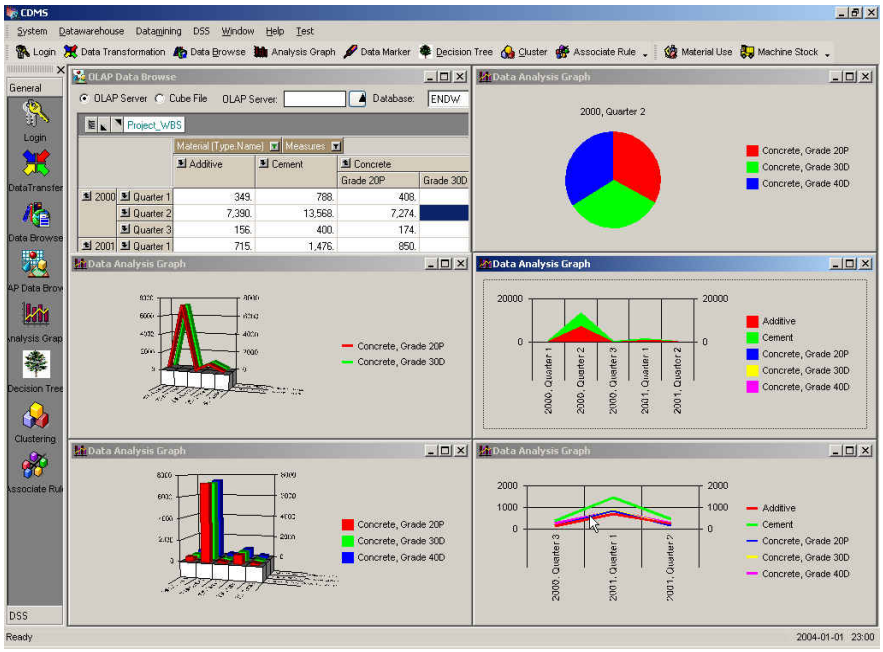

Figure 3 Data browse and graph analysis interface of the DW system

4 DM system for construction management

The establishment of the construction enterprise's DW aims at supporting DM. A series of DM models, facing enterprise's project management, has been built in our system with the historic data in the construction enterprise DW. The tools of client can visualize the DM models and predict useful data in order to assist enterprise's decision-making during projects management.

A typical structure of a DM system based on DW is shown in Figure 4 (Han, 2001):

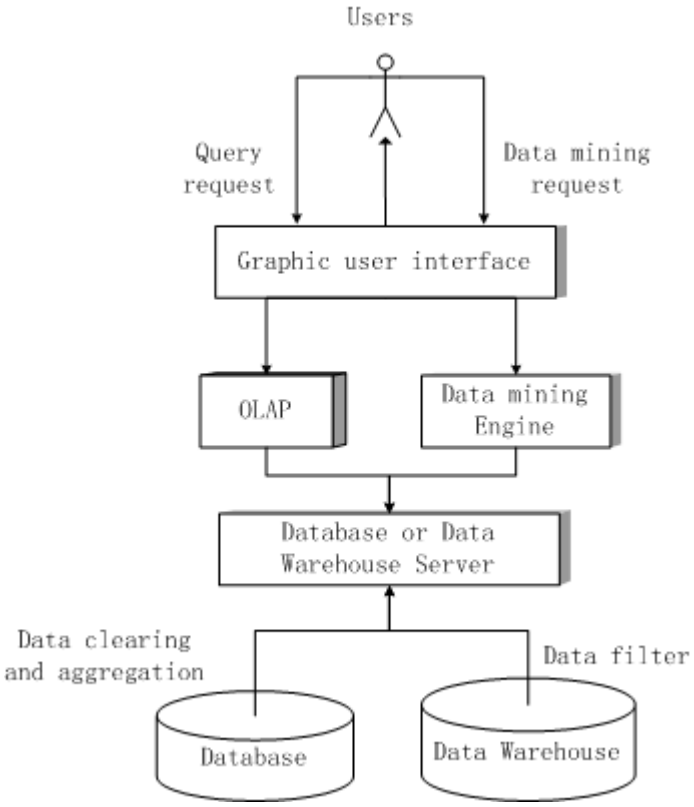

Figure 4 A typical structure of a DM system based on DW

Our DM system mainly deals with two parts of Figure 4: DM Engine and Graphic User Interface.

4.1 DM Engine

In the DM structure of Figure 4, OLAP is in charge of data service, supported by Analysis Services, and DM engine drives DM. In our system, the DM engine contains two core components. One is coming from DM arithmetic offered by Analysis Services, and the other is built with our third-party arithmetic. In our system, three arithmetics have been used: classification of decision tree, cluster analysis and association rule analysis.

4.1.1 Decision Tree

Decision tree is a typical classification. Classification finds the models (or functions) which describe and distinguish data classes and concepts (Han, 2001). A decision tree's inner nodes are partitioned with a selected attribute according to some determined rules. Each branches of one node is a part of the partitions. The leaf nodes of a decision tree represent a distribution. As a familiar method of DM, the characteristic is that it allows users processing a prediction query. The creation process of a decision tree is shown in Figure 5.

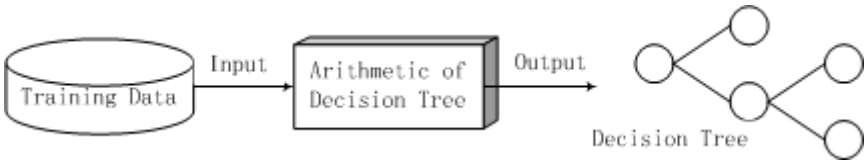

Figure 5 Creation process of a decision tree

There are two steps in the arithmetic of creating decision tree:

(1) The creation of a decision tree

At the beginning, all data are located on the root node. Through a recursion process, arithmetic slice one node's data, according to one of the data's attributes, which will the most information gain when making classification with these data. Do this step recursively, until the tree node can't be sliced. This node is called Leaf Node of a tree. The decision tree arithmetic in our system is basic Avarice Arithmetic. Its process is shown as follows (Han, 2001):

```
Sub BuildTree (S, A)    'S: training data set, A: attribute set for classification
    Build node N using S
    if A is NULL then return N, and mark N as the universal class
    if N is absolutely pure then return N
    Else
        For Each attribute In A
            Calculate the information gain of this node in A
            Find the best attribute as A*, and slice S into Si as the braches of N
            For Each Si in S
                If Si is NULL Then return leaf node and mark Si as the universal class
                Else BuildTree(Si, A-A*)    'recursion
            Next
        Next
    End if
End Sub
```

(2) The pruning of a decision tree

By using the basic avarice arithmetic, the system can create a decision tree. Because of the influence of some noise data or exceptional data, the decision may be too denseness to find useful information. Therefore the pruning arithmetic is necessary to cut out the most impossible branches. Pruning arithmetic includes two methods of pre-printing and post-printing. Our system adopted method of Minimum Description Length (MDL).

4.1.2 Cluster Analysis

Cluster analyzes the data object, but doesn't consider the known tags of classes. It clusters or groups data following the rules which ensure that there is maximal comparability in a class and minimal comparability between the classes. Each cluster should be looked upon as a class object. These clusters can gather those similar events together. Our system adopted one kind of partitioning method: K-Average method, and expanded its validity.

4.1.3 Association Rule

The association analysis discovers association rules. These rules lay out the conditions that attribute-value as a frequent itemset appears in a specific data set frequently. Its excellences are that it can bring clear and useful results, support indirect DM, deal with length-changing data and forecast the expenditure of calculation. But its faults are that the amount of calculations will improve very fast while the issue is becoming larger and larger. Besides this, the rare data is often skipped. The typical arithmetic of Apriori is adopted in our system.

Using these three function models, our system can meet the requirements of processing data in the management of construction projects and assist users' decisions.

4.2 DM Models

In DM environment, model is referred to as a data structure that has stored the instances dealt with DM arithmetic. Our system has a series of typical DM models in server, facing to the requirements of project management in a construction enterprise. Using these built-in models, users can process queries, predictions and decisions expediently. Otherwise, users can establish their own DM models, by provide a data source and appointing a DM arithmetic. The DM models in our system are saved as PMML files. PMML files, based on XML, can be passed and shared expediently through Internet. Users can save DM models in SQL Server or in their local machines in order to advantage the management and application of the models. Part of a PMML file is shown as Figure 6. This file described the decision tree mining model of *Material Use*.

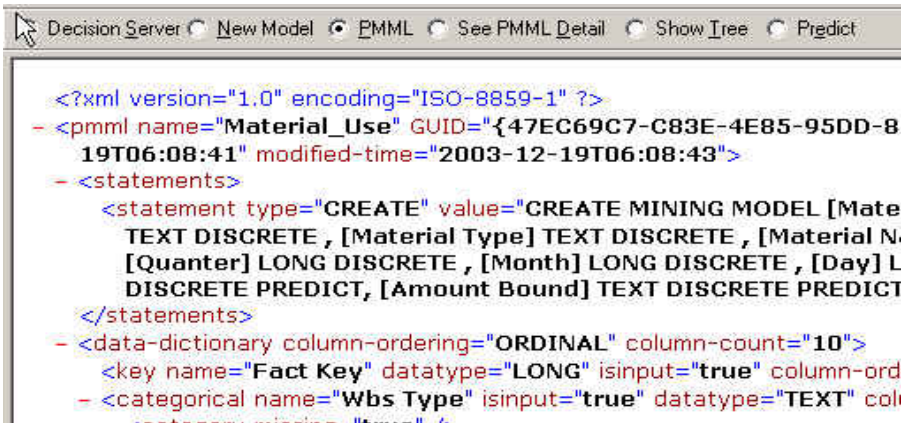

Figure 6 Part of a PMML file of a DM model

4.3 The Graphic User Interface in DM

Graphic user interface of a DM system shows the results of mining to the users in an easy-understanding way. At the same time, users can manage DM models expediently. For example, with the decision-tree model of Material Use, users can find out the distribution of eigenvalues in each key node, so that they can judge the instance of material use in the future. Users also can use the prediction interface by submitting a prediction query and determining some influencing factors, for example material type, structure type, location, project purpose, construction time, etc, in process of material use. Users can gain the predicted values of material use at the same instance. These predicted values can help users to make material plans.

The graphic interface is show as Figure 7, with a decision tree model of Material Use.

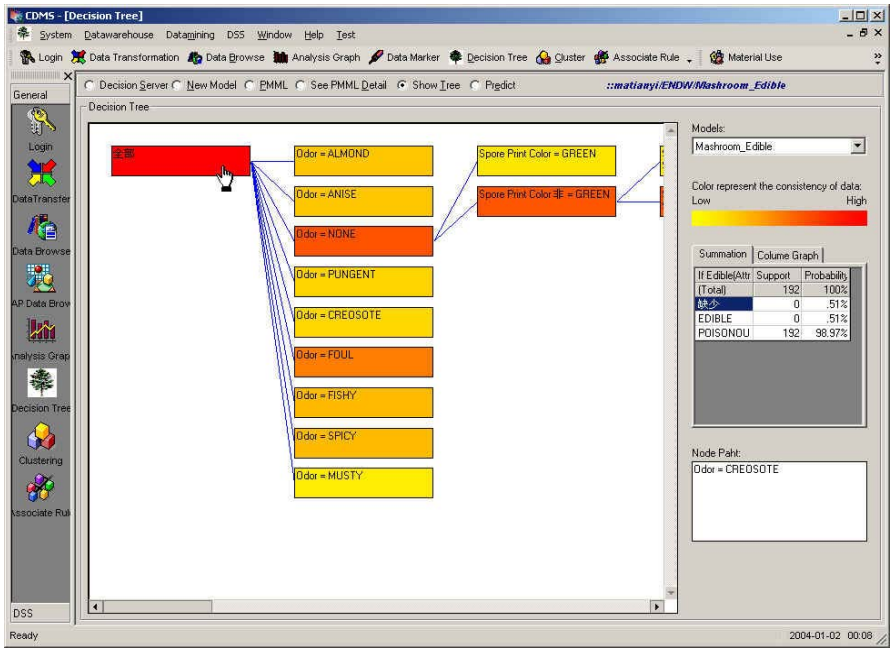

Figure 7 The fraphic interface in the proposed DM system

Once the DW and DM system is uploaded to the enterprise's server, users can browse the information of DM models and submit prediction queries through the Internet.

4.4 Application of the DM System

Our system has developed several specific function modules, including material, machine, human resource, schedule, quality, safety, they are often used during construction management, in order to make scientific decisions and evaluates scheming projects' feasibility.

4.4.1 Analysis of Resources

Here, the resource of construction projects includes humans, materials, machines, and etc, which are necessary in the process of a construction project. First, user should define some specific parameters of a new project, for example the purpose of project, structure type, which country and city, fitment level, time in rain season, total structural areas, etc. By using the DM models and the data of those similar historic projects in DW, user can gain the resources' amount of expenditure (per m2) predicted by DM models. These data can be used to make resource plan of a new project. Of course, user can also look about the correlative information of historic projects. These data or information can be used to form multiform graphs for analysis and decision. The interface of analysis and prediction for material expenditure is shown as Figure 8.

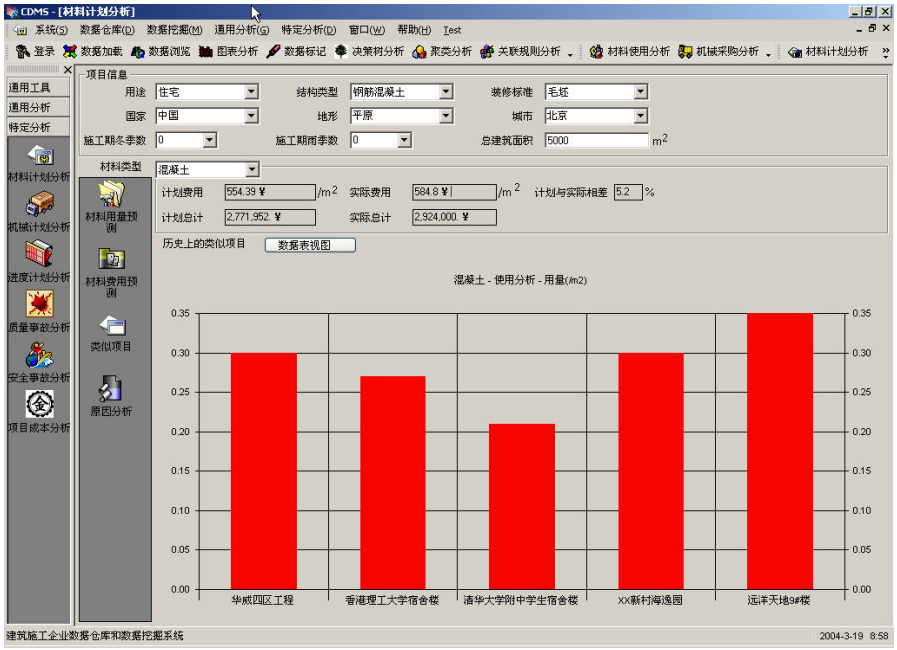

Figure 8 The interface of anlaysis and prediction for material expanditure

4.4.2 Analysis of Construction Schedule

The analysis of construction schedule including two aspects: schedule of the whole project and schedule of WBS. They respectively make use of detail data and aggregated data in DW. The analysis of project just likes the analysis of resource expenditure. But the quality events and safety events will badly influence construction schedule of project. In this instance, users can analysis the schedule of WBS to find the influence of events. The interface of this module is shown as Figure 9. It shows the detail data of WBS, and the association of quality events and the node schedule of WBS.

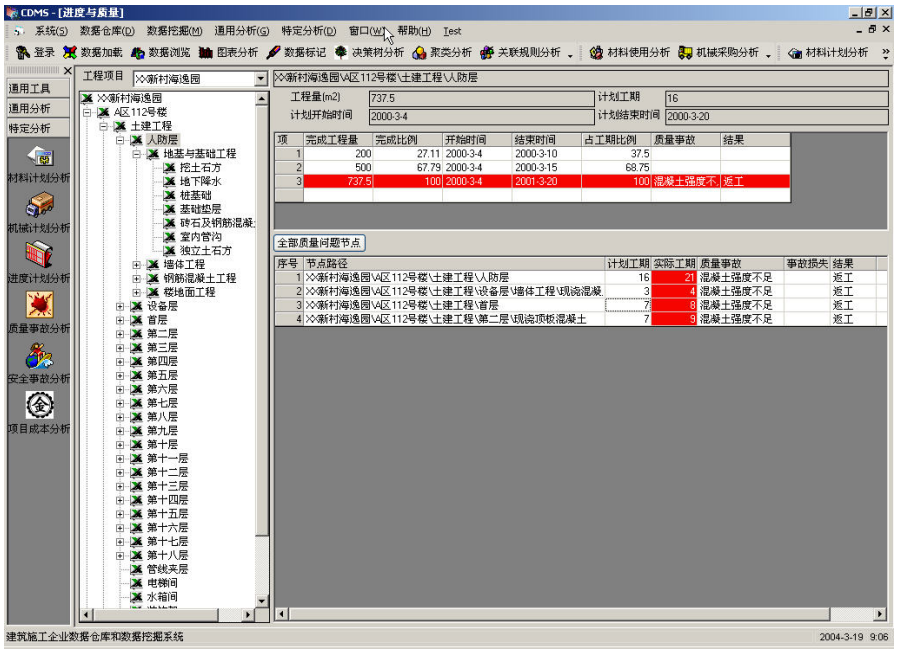

Figure 9 The interface of analyzing construction schedule and quality events

4.4.3 Analysis of Quality and Safety in Construction

According to a new project's characters, such as structure type, node type of WBS, and event level, the system can analyze the events which frequently happened during project construction. The results of the analysis are shown in a grid, and users can discover useful information or use graphic tools to form their own analysis report. Our system also discovers the most influencing factors which influence a project's quality and safety, and the relationship between the events of quality and safety and the project's resource expenditure and progress plan, by using cluster analysis models. The system's interface is shown as Figure 10.

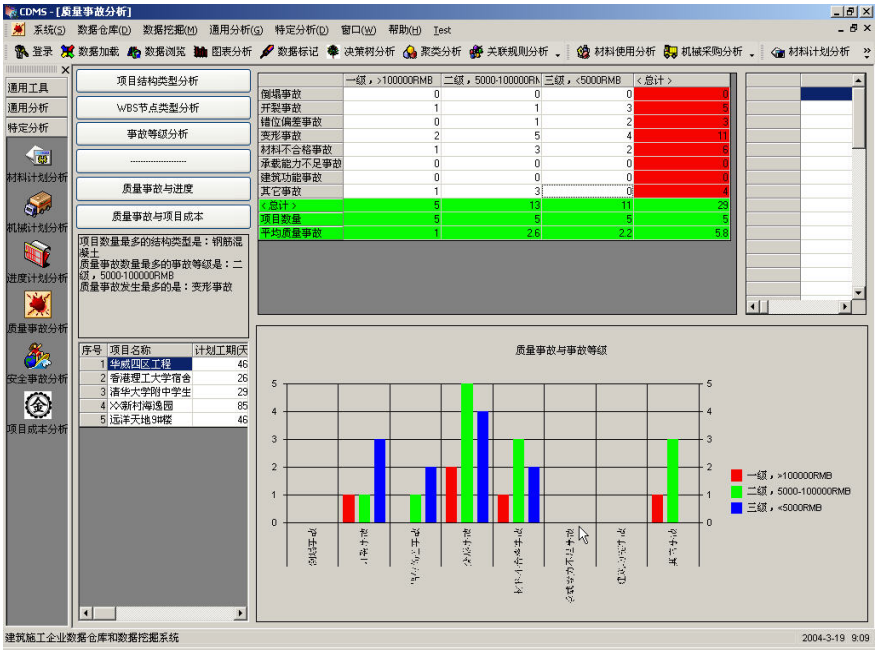

Figure 10 the interface of anlyzing project's quality events

5 Conclusions

This paper introduces the development of a DW system and DM system for construction enterprises in project management. At the same time, facing to the requirements of construction enterprises, we have developed several special analysis tools for resources, schedule, quality, safety and etc. By using this system in several construction projects, the results proved that this system is useful to construction enterprises making decisions during managing projects.

Acknowledgment

This research is supported by the fund from the Tenth five-year Plan administered by The Ministry of Science and Technology of China (2002BA107B), and the **C**enter for **I**nformation **T**echnology in **C**onstruction between Tsinghua University and the Hong Kong Polytechnic University.

References:

Han Jiawei and Kamber Micheline (2001), Data Mining Concepts and Techniques, China Machine Press.

Chau K.W., Ying Cao (2002), The Application of data warehouse and decision support system in construction management, Automation in Construction 12 (2002), 213-224.

J.P.ZHANG, Hong-Jun WANG (2002), Towards 4D Management for Construction Planning and Resource Utilization, The 9th International Conference on Computing in Civil and Building Engineering, Taiwan, pp1281-1286, 2002.4

Hyperion Software Corp. (1999), The Role of OLAP Server in a Data Warehousing Solution.

Inmon W.H. (2000), Building the Data Warehouse-Second Edition, China Machine Press.

Seidman Claude (2002), Data Mining with Microsoft SQL Server 2000 Technical Reference.

Shen Z.Y. (2001), Solution of SQL Server 2000 OLAP – Data Warehouse and Analysis Services, Tsinghua University Press.